

Two-stage residual inclusion estimation: A practitioners guide to Stata implementation

Joseph V. Terza
Department of Economics
Indiana University–Purdue University Indianapolis
Indianapolis, IN
jvterza@iupui.edu

Abstract. Empirical econometric research often requires implementation of nonlinear models whose regressors include one or more endogenous variables—regressors that are correlated with the unobserved random component of the model. In such cases, conventional regression methods that ignore endogeneity will likely produce biased results that are not causally interpretable. Terza, Basu, and Rathouz (2008, *Journal of Health Economics* 27: 531–543) discuss a relatively simple estimation method (two-stage residual inclusion) that avoids endogeneity bias, is applicable in many nonlinear regression contexts, and can easily be implemented in Stata. In this article, I offer a step-by-step protocol to implement the two-stage residual inclusion method in Stata. I illustrate this protocol in the context of a real-data example. I also discuss other examples and pertinent Stata code.

Keywords: st0505, two-stage residual inclusion, endogeneity

1 Introduction

My objective is to develop a simple but consistent estimation protocol in Stata for the parameters of a generic nonlinear regression model with dependent variable Y , which has a vector of independent variables that includes X_u , an unobservable regressor; X_o , a vector of observable regressors that are not correlated with X_u ; and X_e , an observable regressor that is correlated with X_u —that is, X_e is endogenous.¹ The endogeneity of X_e (that is, the correlation between X_e and X_u) confounds the identification and estimation of the possible causal effect of X_e (or any of the other regressors in the model for that matter) on Y . If, for instance, the presence of X_u is ignored, and a conventional regression method is applied, then the corresponding estimate of the effect of X_e will likely be biased, because it will reflect influence that should instead have been attributed to the unobservables. The general modeling and estimation framework discussed by Terza, Basu, and Rathouz (2008) is designed to control for endogeneity, thereby eliminating consequent bias. Their generic model consists of a regression equation with a dependent variable that is the outcome of interest (the outcome equation) and an auxiliary equation that formalizes the correlation between X_e and X_u . The outcome and auxiliary (O&A) equations can each be defined based on either a mini-

1. X_e and X_u may be made up of more than one regressor. We portray them as being single regressors here to simplify exposition.

mally parametric (MP) or a fully parametric (FP) regression structure. Formally, one can specify the outcome component of the model as either

$$Y = \mu(X_e, X_o, X_u; \beta) + e \quad (\text{MP specification}) \quad (1)$$

or

$$f(Y|X_e, X_o, X_u; \beta) \quad (\text{FP specification}) \quad (2)$$

where $\mu(X_e, X_o, X_u; \beta)$ denotes the conditional mean of Y given X_e , X_o , and X_u ; β is a vector of parameters; and $f(Y|X_e, X_o, X_u; \beta)$ is the conditional probability density function of Y given X_e , X_o , and X_u . Similarly, for the auxiliary component of the model, one can posit either

$$X_e = r(W; \alpha) + X_u \quad (\text{MP specification}) \quad (3)$$

or

$$g(X_e|W; \alpha) \quad (\text{FP specification}) \quad (4)$$

where α is a vector of parameters, $r(W; \alpha)$ denotes the conditional mean of X_e given $W = [X_o \ W^+]$, W^+ is a vector identifying instrumental variables, and $g(X_e|W; \alpha)$ is the conditional probability density function of X_e given W . By definition, the elements of W^+ must satisfy the following three conditions: 1) they are correlated with neither X_u nor e ; 2) they can be legitimately excluded from the outcome regression (1); and 3) they are strongly correlated with X_e . Equation (3) [or (4)] formalizes the correlation between X_u and X_e . The correlation between X_u and Y is formalized in the outcome regression (1) [or (2)]. The general two-stage residual inclusion (2SRI) protocol is the following:

First Stage: Apply the appropriate nonlinear least squares (NLS) [maximum likelihood (ML)] estimator to (3) [or (4)] to consistently estimate α .² The residuals from this regression are

$$\hat{X}_u = X_e - r(W; \hat{\alpha}) \quad (5)$$

where $\hat{\alpha}$ denotes the first-stage consistent estimate of α . Note that the FP specification in (4) will always imply the existence of a regression specification akin to (3), from which residuals, as defined in (5), can be obtained. To complete the first stage of 2SRI, save the residuals defined in (5).

Second Stage: To consistently estimate β , apply the appropriate NLS [ML] estimator to (1) [or (2)], with X_u replaced by \hat{X}_u .³

Note that one can use any combination of MP or FP specifications for the first and second stages of the 2SRI model. Correspondingly, any combination of NLS or ML can be implemented for first- and second-stage estimation. In the majority of applied

2. The first-stage ML estimator is the maximizer of $\sum_{i=1}^n \ln\{g(X_{ei}|W_i; \alpha)\}$ with respect to α , where X_{ei} and W_i denote the observed values of X_e and W for the i th observation in the sample and $i = 1, \dots, n$.

3. The second-stage ML estimator is the maximizer of $\sum_{i=1}^n \ln\{f(Y_i|X_{ei}, X_{oi}, \hat{X}_{ui}; \beta)\}$ with respect to β , where Y_i and X_{oi} denote the observed values of Y and X_o for the i th observation in the sample and where \hat{X}_{ui} is the first-stage residual for the i th observation in the sample.

settings, the 2SRI estimates of α and β are easy to obtain via packaged Stata commands. The asymptotically correct standard errors (ACSE), for use in estimation of confidence intervals and t statistics for testing hypotheses about the elements of β , can be calculated with additional Mata commands.

Before moving on to an example, note that the above model specification and corresponding estimator do not necessarily constitute a control function method (CFM) as defined by [Blundell and Powell \(2003\)](#).⁴ The assumption that I maintain above is that the O&A regressions are correctly specified by the researcher. As [Terza, Basu, and Rathouz \(2008\)](#) show, under this assumption, the 2SRI estimator consistently estimates the model parameters.⁵ To qualify as a CFM with accompanying consistency and robustness properties, the above 2SRI approach must satisfy other conditions. For a detailed discussion of such conditions, see [Wooldridge \(2014, 2015\)](#). To maintain the focus of this article (imparting practical aspects of 2SRI implementation in Stata), we abstract from such issues in the following sections. For simplicity of illustration and didactics, we maintain that for a correctly specified model, 2SRI affords the applied researcher a consistent, coherent but simple way to do empirical analyses for a very general class of nonlinear data-generating processes.

Consider the regression model of [Mullahy \(1997\)](#), in which the objective is to draw causal inferences regarding the effect of prenatal smoking (X_e -CIGSPREG) on infant birthweight (Y -BIRTHWTLB) while controlling for infant birth order (PARITY), race (WHITE), and sex (MALE). The regression model for the birthweight outcome that he proposed can be written in the MP form⁶

$$Y = \exp(X_e\beta_e + X_o\beta_o + X_u\beta_u) + e \quad (\text{outcome regression}) \quad (6)$$

where X_u comprises unobservable variables that are potentially correlated with prenatal smoking (for example, general “health mindedness” of the mother), e is the regression error term, $X_o = [\text{PARITY WHITE MALE}]$ is a row vector of regressors that are uncorrelated with X_u , and e and the β ’s are the regression parameters. At issue here is the fact that there exist unobservables (as captured by X_u) that are correlated with both Y and X_e . In other words, X_e is endogenous. For illustrative purposes, we specify an FP version of the auxiliary component of the model in which

$$g(X_e|W; \alpha^*) = \{1 - \Phi(W\alpha_1)\}^{I(X_e=0)} \times \{\Phi(W\alpha_1) \ln \varphi(X_e, W\alpha_2^*, \sigma^2)\}^{\{1-I(X_e=0)\}} \quad (7)$$

-
4. Under the assumptions of [Blundell and Powell \(2003\)](#) (mainly linearity), in their discussion of CFM, the condition in expression (63) of [Wooldridge \(2014\)](#) is implied. Wooldridge also notes that, although (63) is not precluded in the nonlinear 2SRI framework, it is also not implied. Therefore, (63) must be imposed if 2SRI is to be interpreted as a CFM as in [Blundell and Powell \(2003\)](#).
 5. Under the assumption that the model is correctly specified (and other general conditions), the consistency of the 2SRI estimator follows from the fact that it is a member of the class of two-stage M -estimators (see [Newey and McFadden \[1994, sec. 6\]](#); [White \[1994, chap. 6\]](#); [Wooldridge \[2010, chap. 12\]](#)).
 6. [Mullahy \(1997\)](#) does not explicitly specify the model in terms of the unobservable X_u . Nevertheless, (6) is substantively identical to Mullahy’s (1997) model (see [Terza \[2006\]](#)).

where $\alpha^{*'} = [\alpha_1' \ \alpha_2^{*'}]$, $\ln \varphi(A, b, c)$ denotes the probability density function of the log-normal random variable A with central tendency parameter b and dispersion parameter c , $W = [X_o \ W^+]$, and $W^+ = [\text{EDFATHER} \ \text{EDMOTHER} \ \text{FAMINCOME} \ \text{CIGTAX}]$, with

EDFATHER = paternal schooling in years
EDMOTHER = maternal schooling in years
FAMINCOME = family income

and

CIGTAX = cigarette tax

The specification in (7) indicates that prenatal smoking follows a two-part model with a probit formulation for the extensive margin (EM) and a lognormal intensive margin (IM). This is, in fact, a reasonable specification because a) there is a substantial proportion of nonsmokers in the population (and sample) of pregnant women; and b) the decision to smoke or not probably differs systematically from the decision regarding how much to smoke (among those who have decided to smoke at all). Based on (7), we can write the auxiliary regression as

$$X_p = \Phi(W\alpha_1)\exp(W\alpha_2) + X_u \quad (\text{auxiliary regression}) \quad (8)$$

where α_2 is the same as α_2^* , with the constant term shifted by $+(\sigma^2/2)$, because (7) implies that $E[X_e|W] = \Phi(W\alpha_1)\exp\{W\alpha_2^* + (\sigma^2/2)\}$. From (8), we have that $r(W; \alpha) = \Phi(W\alpha_1)\exp(W\alpha_2)$ and

$$X_u = X_e - \Phi(W\alpha_1)\exp(W\alpha_2) \quad (9)$$

where $\alpha' = [\alpha_1' \ \alpha_2']$. In the sequel, we will refer to the model in (6) through (9) as *the example*. For the generic nonlinear model with endogeneity [(1) through (4)], we offer a step-by-step protocol for using Stata and Mata to obtain the 2SRI estimate of β and the corresponding ACSE.⁷ We use the example to illustrate each of the steps.

2 The step-by-step 2SRI protocol

In detailing this protocol, we assume that the data have been input and that the analysis sample comprises n observations on the following variables: Y , X_e , X_o , and W_{plus} , corresponding to Y , X_e , X_o , and W^+ as generically defined above.

7. There are two other ways to calculate the standard errors: bootstrapping and the resampling method proposed by Krinsky and Robb (1986, 1990). For detailed discussions and pro-and-con evaluations of the bootstrapping and Krinsky and Robb (1986, 1990) methods, see Dowd, Greene, and Norton (2014). Dowd, Greene, and Norton (2014) also discuss the ACSE approach, but the formulation they offer [in particular, (17)] is based on an assumption that is usually invalid in econometric applications. See Terza (2016b) for details.

Step a: Specify the O&A components of the 2SRI model.

Any of four O&A combinations is possible based on the choice of MP versus FP specifications for each of the two estimation stages. For the second-stage outcome component, one can use $\mu(X_e, X_o, X_u; \beta)$ [MP specification in (1)] or $f(Y|X_e, X_o, X_u; \beta)$ [FP specification in (2)]. To make their dependence on α and β explicit, and for convenience of exposition, we rewrite the MP and FP versions of the outcome regression, respectively, as

$$\mu^*(X_e, W; \alpha, \beta) = \mu[X_e, X_o, \{X_e - r(W; \alpha)\}; \beta] \quad (10)$$

and

$$f^*(Y|X_e, W; \alpha, \beta) = f[Y|X_e, X_o, \{X_e - r(W; \alpha)\}; \beta]$$

For the first-stage auxiliary component, one can use $r(W; \alpha)$ [MP specification in (3)] or $g(X_e|W; \alpha)$ [FP specification in (4)]. MP (FP) O&A 2SRI components can be estimated via NLS (ML). In the example using (9), the following version of (10) is relevant,

$$\mu^*(X_e, W; \alpha, \beta) = \exp[X_e \beta_p + X_o \beta_o + \{X_e - \Phi(W \alpha_1) \exp(W \alpha_2)\} \beta_u] \quad (11)$$

where $\beta' = [\beta_e \ \beta_o \ \beta_u]$.

Step b: Derive the requisite analytic components for calculation of the ACSE.

As Terza (2016a) shows, the exact form of the ACSE depends on the estimation method used in the second stage of 2SRI—NLS (for the MP specification) versus a maximum likelihood estimator (MLE) (for the FP specification). When an MLE is used in the second stage, the ACSE for the k th element of β is the square root of the k th diagonal element of the matrix,

$$V(\hat{\beta}) A V(\hat{\alpha}) A' V(\hat{\beta}) + V(\hat{\beta}) \quad (12)$$

where $V(\hat{\alpha})$ and $V(\hat{\beta})$ are the estimates of the covariance matrices output by the relevant Stata commands for the first and second stages of 2SRI, respectively, and

$$A = \sum_{i=1}^n \nabla_{\beta} \ln \hat{f}_i^* \nabla_{\alpha} \ln \hat{f}_i^* \quad (13)$$

with $\nabla_c \ln \hat{f}_i^*$ defined as the gradient of $f^*(Y|X_e, W; \alpha, \beta)$ with respect to c ($c = \alpha$ or β) evaluated at X_{ei} , $W_i[X_{oi} \ W_i^+]$, $\hat{\alpha}$, and $\hat{\beta}$ (“ i ” denotes the i th observation in the sample; $i = 1, \dots, n$). In this case, analytic expressions for $\nabla_{\beta} \ln f^*$ and $\nabla_{\alpha} \ln f^*$ must be derived.

Similarly, Terza (2016a) shows that when NLS is used in the second stage, the ACSE for the k th element of β is the square root of the k th diagonal element of the matrix,

$$B_1^{-1} B_2 V(\hat{\alpha}) B_2' B_1^{-1} + V(\hat{\beta}) \quad (14)$$

where $V(\hat{\alpha})$ and $V(\hat{\beta})$ are the estimated variance–covariance matrices of the first- and second-stage estimators of α and β , respectively, as output by Stata,

$$B_1 = \sum_{i=1}^n \nabla_{\beta} \hat{\mu}_i^{*'} \nabla_{\beta} \hat{\mu}_i^* \quad (15)$$

and

$$B_2 = \sum_{i=1}^n \nabla_{\beta} \hat{\mu}_i^{*'} \nabla_{\alpha} \hat{\mu}_i^* \quad (16)$$

with $\nabla_c \hat{\mu}_i^*$ defined as the gradient of $\mu^*(X_e, W; \alpha, \beta)$ with respect to c ($c = \alpha$ or β) evaluated at X_{ei} , $W_i = [X_{oi} \ W_i^+]$, $\hat{\alpha}$, and $\hat{\beta}$. This step requires that the user supply analytic expressions for $\nabla_{\beta} \mu^*$ and $\nabla_{\alpha} \mu^*$. In the example, it follows from (3) that

$$\nabla_{\beta} \mu^* = \exp(X\beta)X$$

and

$$\nabla_{\alpha} \mu^* = [\nabla_{\alpha_1} \mu^* \quad \nabla_{\alpha_2} \mu^*]$$

where

$$\begin{aligned} \nabla_{\alpha_1} \mu^* &= -\beta_u \exp(X\beta) \exp(W\alpha_2) \varphi(W\alpha_1) W \\ \nabla_{\alpha_2} \mu^* &= -\beta_u \exp(X\beta) \exp(W\alpha_2) \Phi(W\alpha_1) W \\ X &= [X_e \ X_o \ X_u] \quad \text{and} \quad W = [X_o \ W^+] \end{aligned}$$

Therefore,

$$\nabla_{\beta} \hat{\mu}_i^* = \exp(X_i \hat{\beta}) X_i$$

and

$$\nabla_{\alpha} \hat{\mu}_i^* = [\nabla_{\alpha_1} \hat{\mu}_i^* \quad \nabla_{\alpha_2} \hat{\mu}_i^*]$$

where

$$\begin{aligned} \nabla_{\alpha_1} \mu^* &= -\hat{\beta}_u \exp(X_i \hat{\beta}) \exp(W_i \hat{\alpha}_2) \varphi(W_i \hat{\alpha}_1) W_i \\ \nabla_{\alpha_2} \mu^* &= -\hat{\beta}_u \exp(X_i \hat{\beta}) \exp(W_i \hat{\alpha}_2) \Phi(W_i \hat{\alpha}_1) W_i \\ X_i &= [X_{ei} \ X_{oi} \ \hat{X}_{ui}] \quad \text{and} \quad \hat{X}_{ui} = X_{ei} - \Phi(W_i \hat{\alpha}_1) \exp(W_i \hat{\alpha}_2) \end{aligned}$$

Generally (second-stage ML or NLS), based on standard asymptotic theory, the “ t statistic” is

$$\frac{\hat{\beta}(k) - \beta(k)}{\sqrt{\hat{D}(k)}} \quad (17)$$

for the k th element of β , and $[\beta(k)]$ is asymptotically standard normally distributed, where $\hat{\beta}(k)$ is the 2SRI estimator of $\beta(k)$ and $\hat{D}(k)$ denotes the k th diagonal element of (3) or (13). This t statistic can be used to test the hypothesis that $\beta(k) = \beta(k)^0$ for $\beta(k)^0$ —a given null value of $\beta(k)$.

Step c: Apply the appropriate Stata commands for $r(W, \alpha)$ $[g(X_e|W; \alpha)]$ when the first stage is NLS [MLE] to obtain the first-stage estimate of α by regressing X_e on X_o and W plus.

In the example, the parameter vector for the first part (EM) of the auxiliary component of the model (α_1) can be estimated by applying the Stata `probit` command to the full sample, with $[1 - I(X_e = 0)]$ as the dependent variable and W as the vector of regressors, where $I(C)$ denotes the indicator function that takes the value 1 if condition C holds and 0 otherwise. The parameters of the second part (IM) of the auxiliary component of the model (α_2) can be consistently estimated by applying the Stata `glm` command to the subsample of nonzero smokers, with X_e as the dependent variable and W as the vector of regressors.

```

/*****
** Generate the binary smoking variable.      **
*****/
gen ANYCIGS=CIGSPREG>0

/*****
** 2SRI first-stage first-part probit estimates.**
*****/
/*Step c*/
probit ANYCIGS PARITY WHITE MALE EDFATHER EDMOTHER ///
      FAMINCOM CIGTAX88
.
.
.

/*****
** 2SRI first-stage second-part probit NLS      **
** estimates.                                  **
*****/
/*Step c*/
glm CIGSPREG PARITY WHITE MALE EDFATHER EDMOTHER ///
    FAMINCOM CIGTAX88 if ANYCIGS==1,      ///
    family(gaussian) link(log) vce(robust)

```

Step d: Use the appropriate command or option to calculate and save the first-stage regression residuals, say, as the additional variable X_{uhat} .

In the context of the example, we have

```

/*Step d*/
predict CIGPROB
.
.
.

/*Step d*/
predict CIGMEAN
.
.
.

```

```

/*****
** Generate the first-stage residuals.      **
*****/
/*Step d*/
gen Xuhat=CIGSPREG-CIGPROB*CIGMEAN

```

The first (second) `predict` is placed immediately after the `probit (glm)` command in step c and produces the first-stage first (or second)-part `probit` (exponential regression) predictions $\Phi(W_i\hat{\alpha}_1)$ [$\exp(W_i\hat{\alpha}_2)$].

Step e: Use the appropriate Stata and Mata commands to save the vector of first-stage coefficient estimates and its corresponding estimated covariance matrix (as calculated and output by the relevant Stata commands used in step c) so that they are accessible in Mata; call them, for example, `alpha1hat` and `Valpha1hat`, respectively.

In the context of the example, we have

```

/*****
** Save the first-stage first-part probit      **
** estimates and estimated covariance matrix.  **
*****/
/*Step e*/
mata: alpha1hat=st_matrix("e(b)")`
mata: Valpha1hat=st_matrix("e(V)")
.
.
.
/*****
** Save the first-stage second-part NLS      **
** estimates and estimated covariance matrix.  **
*****/
/*Step e*/
mata: alpha2hat=st_matrix("e(b)")`
mata: Valpha2hat=st_matrix("e(V)")

```

The first (second) pair of Mata commands is placed immediately after the `predict CIGPROB (predict CIGMEAN)` command in step d. The `st_matrix(name)` function turns the Stata matrix *name* into a Mata matrix. In this context, the `probit` and `glm` commands produce the vector of coefficient parameter estimates `e(b)` and estimated covariance matrix `e(V)` among their stored results. The `st_matrix()` function transforms them to Mata-usable format.

Step f: Apply the appropriate Stata commands for $\mu(X_e, X_o, X_u; \beta)$ [$f(Y|X_e, X_o, X_u; \beta)$] when the 2SRI second stage is NLS [ML] to obtain the second-stage estimate of β by regressing `Y` on `Xe`, `Xo`, and `Xuhat`.

In the context of the example, we have

```

/*Step f*/
glm BIRTHWTLB CIGSPREG PARITY WHITE MALE Xuhat, ///
    family(gaussian) link(log) vce(robust)

```


Step g: Use the Stata and Mata commands to save the vector of second-stage coefficient estimates and its corresponding estimated covariance matrix (as calculated and output by the relevant Stata commands used in step f) so that they are accessible in Mata; call them, for example, `betahat` and `Vbetahat`, respectively (you might also have to single out $\hat{\beta}_u$).

In the context of the example, we have

```
/*Step g*/
mata: betahat=st_matrix("e(b)")`
mata: Vbetahat=st_matrix("e(V)")
mata: Bu=betahat[5]
```

The last statement uses matrix subscripting and the fact that $\hat{\beta}_u$ is the fifth element of the estimated coefficients of the exponential outcome regression.

Step h: Construct `X` and `W` matrices in Mata, where `X` is the matrix that has columns that are `Xe`, `Xo`, and a constant term (a column vector of 1s); and `W` has columns `Xo`, `Wplus`, and a constant term.⁸

In the context of the example, we have

```
/*Step h*/
putmata BIRTHWTLB CIGSPREG ANYCIGS PARITY WHITE ///
        MALE EDFATHER EDMOTHER FAMINCOM CIGTAX88 Xuhat

/*Step h*/
mata: X=CIGSPREG, PARITY, WHITE, MALE,          ///
        Xuhat, J(rows(PARITY),1,1)
mata: W=PARITY, WHITE, MALE, EDFATHER, EDMOTHER, ///
        FAMINCOM, CIGTAX88, J(rows(PARITY),1,1)
```

The `putmata` command converts designated variables in the relevant Stata dataset to vectors in Mata-usable format.

Step i: Use `alphahat`, `betahat`, `X`, `W`, and the analytic results obtained in step b to construct the two gradient matrices needed to calculate the correct standard errors for `betahat`, say, `gradbeta` and `gradalpha`. Note that `gradbeta` will have `n` rows and `K` columns, where `K` is the column dimension of `X`, and `gradalpha` will have `n` rows and `S` columns, where `S` is the column dimension of `W`. The exact forms of these gradient matrices will depend on whether ML or NLS was implemented in the second stage of the 2SRI estimator. If ML was used, then the *i*th rows of `gradbeta` and `gradalpha` will be $\nabla_{\beta} \ln \hat{f}_i^*$ and $\nabla_{\alpha} \ln \hat{f}_i^*$, respectively, as defined in (12). If the 2SRI second stage is NLS, then the *i*th rows of `gradbeta` and `gradalpha` will be $\nabla_{\beta} \hat{\mu}_i^*$ and $\nabla_{\alpha} \hat{\mu}_i^*$, respectively, as defined in (14) and (16).

8. Be sure that the ordering of the columns of `X` and `W` (including the constant term) conforms to the ordering of the estimated coefficients in `betahat` and `alphahat`.

In the context of the example, we have

```
/*Step i*/
mata: gradbeta=exp(X*betahat):*X
mata: gradalpha1=
-Bu:*exp(X*betahat):*normalden(W*alpha1hat):*exp(W*alpha2hat):*W
mata: gradalpha2=
-Bu:*exp(X*betahat):*normal(W*alpha1hat):*exp(W*alpha2hat):*W
mata: gradalpha=gradalpha1,gradalpha2
```

Step j: Calculate \mathbf{A} [\mathbf{B}_1 and \mathbf{B}_2] as defined in (12) [(14) and (16)].

If the 2SRI second stage is ML, then calculate the \mathbf{A} matrix as⁹

```
A = gradbeta' * gradbeta
```

based on (12). Because the 2SRI second stage in the example is NLS, we calculate the \mathbf{B}_1 and \mathbf{B}_2 matrices as

```
/*Step j*/
mata: B1=gradbeta'*gradbeta
mata: B2=gradbeta'*gradalpha
```

based on (14) and (16), respectively.

Step k: Calculate the asymptotic covariance matrix of $\hat{\beta}$.

If the 2SRI second stage is ML, then calculate the estimated asymptotic covariance matrix of `betahat` as

```
AVARBeta = Vbetahat * A * Valphahat * A' Vbetahat' + Vbetahat
```

based on (3). Because the 2SRI second stage in the example is NLS, we calculate the estimated asymptotic covariance matrix of `betahat` as

```
/*Step k*/
mata: Valphahat=blockdiag(Valpha1hat,Valpha2hat)
mata: Dhat=invsym(B1)*B2*Valphahat*B2'*invsym(B1)+Vbetahat
```

based on (13). Note that we first had to stack up the full estimated covariance matrix of $\hat{\alpha} = [\hat{\alpha}'_1 \hat{\alpha}'_2]$ from the first- and second-part outputs for the first-stage 2SRI estimate of α .

9. Here we use the following summation or matrix equality: Let \mathbf{Z}_i and \mathbf{Q}_i be the K and S dimensional row vectors, respectively ($i = 1, \dots, n$), and let \mathbf{Z} and \mathbf{Q} be the $n \times K$ and $n \times S$ matrices with i th rows that are \mathbf{Z}_i and \mathbf{Q}_i , respectively; then

$$\sum_{i=1}^n \mathbf{Z}'_i \mathbf{Q}_i = \mathbf{Z}' \mathbf{Q}$$

Step l: Calculate the vector of asymptotic standard errors for $\hat{\beta}$.

Regardless of the estimator used in the 2SRI second stage, use

```
mata: ACSE=sqrt(diagonal(AVARBeta))
```

Step m: Calculate the vector of asymptotic t statistics to be used to test the conventional null hypothesis regarding the elements of β (namely, $H_0: \beta_k = 0$, where β_k denotes the k th element of β).

Regardless of the estimator used in the 2SRI second stage, use

```
/*Step m*/  
mata: Betatstats=Betahat:/ACSE.
```

The k th element of `Betatstats` corresponds with (17). The full Stata code for this protocol as it pertains to the example is given in the appendix.

I applied the above 2SRI estimation protocol to the same dataset analyzed by Mullahy (1997). The estimation results for α and β are reported in tables 1 and 2, respectively. The correct asymptotic t statistics for the 2SRI estimate of β , reported in column 3 of table 2, were calculated using (13). In table 2, we also display Mullahy's generalized method of moments (GMM) estimates and, as a baseline, report the conventional NLS estimates that ignore potential endogeneity. As an indicator of the strength of the instrumental variables (that is, the elements of W^+), we conducted a Wald test of their joint significance. The value of the chi-squared test statistic is 49.33, so the null hypothesis that their coefficients are jointly zero is roundly rejected at any reasonable level of significance. The second-stage 2SRI estimates shown in table 2 (column 2) are virtually identical to Mullahy's GMM estimates (column 5), but the former, unlike the latter, provide a direct test of the endogeneity of the prenatal smoking variable via the asymptotic t statistic (5th element of $\hat{\beta}$) for the coefficient of $X_u[\hat{\beta}_u = \hat{\beta}(5)]$ with $H_0: \beta_u = \beta(5) = 0$. According to the results of this test, the exogeneity null hypothesis is rejected at nearly the 1% significance level. To get a sense of the bias from neglecting to account for the two-stage nature of the estimator in the calculation of the asymptotic standard errors, in table 2 (last column), we also display the “packaged” second-stage `glm` t statistics as reported in the Stata output. The mean absolute bias across these uncorrected asymptotic t statistics for the four control variables and X_u is nearly 9%.

Table 1. 2SRI first-stage estimates

Variable	Estimate	Asymptotic t statistic	p -value
First-stage estimate of α_1			
PARITY	0.02	0.39	0.696
WHITE	0.25	2.16	0.031
MALE	-0.16	-1.88	0.060
EDFATHER	-0.02	-2.38	0.017
EDMOTHER	-0.12	-5.54	0.000
FAMINCOM	-0.01	-2.87	0.004
CIGTAX	0.01	2.25	0.024
Constant	0.56	1.93	0.054
First-stage estimate of α_2			
PARITY	0.10	1.34	0.182
WHITE	0.00	0.00	0.998
MALE	0.21	2.13	0.033
EDFATHER	-0.02	-1.43	0.153
EDMOTHER	-0.03	-0.87	0.386
FAMINCOM	0.00	0.28	0.778
CIGTAX	0.00	-0.39	0.697
Constant	2.82	6.00	0.000

$n = 1388$

Table 2. 2SRI second-stage, GMM, and NLS estimates

Variable	Estimate	2SRI		GMM			NLS	
		Correct asymptotic t statistic	Uncorrected asymptotic t statistic	Estimate	Asymptotic t statistic	Estimate	Asymptotic t statistic	Estimate
CIGS	-0.01	-4.07	-4.41	-0.01	-3.46	0.00	-5.62	
PARITY	0.02	3.36	3.66	0.02	3.33	0.01	2.99	
WHITE	0.05	4.45	4.61	0.05	4.44	0.06	4.75	
MALE	0.03	2.80	2.90	0.03	2.95	0.03	2.90	
X_u	0.01	2.66	2.89	-	-	-	-	
Constant	1.94	124.67	129.70	1.94	121.71	1.93	133.70	

$n = 1388$

3 Other oft-encountered O&A combinations

Nonlinearity in regression modeling is typically implied by limitations on the support of the dependent variable. For instance, the linear specification is clearly unappealing for models with binary or fractional support. Another commonly encountered dependent variable type that prompts nonlinear modeling is one whose support is the nonnegative half of the real line (including zero). In the previous section, in the context of the example, we discussed a particular version of this case, in which there is i) a nontrivial proportion of zeros in the population (sample); and ii) a reason to believe that the EM (zero or not) should be modeled differently from the IM (value of the dependent variable conditional on it being nonzero). In a simpler (nested) version of this model, there is no need to distinguish between the EM and IM in modeling. In the example, if there were no reason to believe that the decision regarding whether or not to smoke during pregnancy (IM) is systematically different from one's choice of how much to smoke (EM), then we would replace (7) with

$$X_e = \exp(W\alpha) + X_u$$

and implement NLS for 2SRI first-stage estimation of α . We leave it to the reader to supply the details of the above step-by-step 2SRI protocol for this case. In the remainder of this section, we discuss binary and fractional O&A specifications.

Consider the details of the step-by-step protocol when X_e is binary and Y is fractional. From the following discussion of this case, the reader should be able to infer the details of the protocol for the remaining three possible O&A specifications involving these two variable types.

Step a: In this case, the first- and second-stage estimators are ML and NLS, respectively. The conditional pdf for ML in the first stage is

$$g(X_e|W; \alpha) = \Phi(W\alpha)^{X_e} \{1 - \Phi(W\alpha)\}^{1-X_e} \quad (18)$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function.¹⁰ Note that (15) implies that $r(W, \alpha) = \Phi(W\alpha)$. The functional form for the outcome regression in (1) and (10) is

$$\begin{aligned} \mu^*(X_e, W; \alpha, \beta) &= \mu(X_e, X_o, \{X_e - r(W; \alpha)\}; \beta) \\ &= \Phi(X\beta) = \Phi[X_e\beta_p + X_o\beta_o + \{X_e - \Phi(W\alpha)\}\beta_u] \end{aligned}$$

Step b:

$$\nabla_{\beta} \mu^* = \varphi(X\beta)X$$

and

$$\nabla_{\alpha} \mu^* = -\beta_u \varphi(X\beta) \varphi(W\alpha) W$$

10. $\Phi(\cdot)$ can be replaced here by any convenient (packaged) cumulative distribution function.

Step c:

```
/*step c*/  
probit Xe Xo Wplus
```

Step d:

```
/*step d*/  
predict phiWalpha, p  
gen Xuhat=Y-phiWalpha
```

Step e:

```
/*step e*/  
mata: alphahat=st_matrix("e(b)")´  
mata: Valphahat=st_matrix("e(V)")
```

Step f:

```
/*step f*/  
glm Y Xe Xo Xuhat,family(gaussian) link(probit) vce(robust)
```

Step g:

```
/*step g*/  
mata: betahat=st_matrix("e(b)")´  
mata: Vbetahat=st_matrix("e(V)")  
mata: Bu=betahat[3]
```

Step h:

```
/*step h*/  
putmata Y Xe Xo Wplus Xuhat  
mata: X=Xe, Xo, Xuhat, J(rows(Xo),1,1)  
mata: W=Xo, Wplus, J(rows(Xo),1,1)
```

Step i:

```
/*step i*/  
mata: gradbeta=normalden(X*betahat):*X  
mata: gradalpha=-Bu:*normalden(X*betahat):*/*  
*/normalden(W*alphahat):*W
```

Step j:

```
/*step j*/  
mata: B1 = gradbeta´*gradbeta  
mata: B2 = gradbeta´*gradalpha
```

Step k:

```
/*step k*/  
mata: AVARBeta=invsym(B1)*B2*Valphahat*B2´invsym(B1)/*  
*/+ Vbetahat
```

Step l:

```
/*step l*/  
mata: ACSE = sqrt(diagonal(AVARBeta))
```

Step m:

```
/*step m*/  
mata: ACtstats=betahat:/ACSE
```

4 Summary and discussion

I reviewed the 2SRI method for nonlinear models with endogenous regressors and offered a step-by-step protocol for its implementation in Stata. I illustrated its application with real data for when both X_e and Y are nonnegative. In empirical practice, cases in which X_e , Y , or both are binary or fractional often arise. I detailed Stata and Mata implementation of the protocol for the version of the model in which X_e is binary and Y is fractional. I hope that these examples will serve to demonstrate the ease with which the protocol can be extended to models involving other variable-type configurations not explicitly covered here. In particular, the class of nonnegative dependent variables encompasses important subtypes; for example, count variables and continuous variables with support that does not include 0. For instance, one might seek to fit a model with an endogenous count regressor and an outcome whose distribution is skewed with 0 excluded. In this case, $g(X_e|W; \alpha)$ might be specified as Poisson and $f(Y|X_e, W, X_u; \alpha, \beta)$ as generalized Gamma. In Stata, the first-stage MLE of α would be obtained using the `poisson` command. The `streg` command with the `distribution(ggamma)` option would be used to obtain the second-stage MLE of β . The ACSEs for the elements of β would be obtained using our proposed protocol.

5 Acknowledgments

This research was supported by a grant from the Agency for Healthcare Research and Quality (R01 HS017434-01). This article was presented at the Stata Conference in Chicago, IL, July 28–29, 2016. Please do not quote without the author’s permission.

6 References

- Blundell, R., and J. L. Powell. 2003. Endogeneity in nonparametric and semiparametric regression models. In *Advances in Economics and Econometrics: Theory and Applications, Eight World Congress*, vol. 2, ed. M. Dewatripont, L. P. Hansen, and S. J. Turnovsky, 312–357. Cambridge: Cambridge University Press.
- Dowd, B. E., W. H. Greene, and E. C. Norton. 2014. Computation of standard errors. *Health Services Research* 49: 731–750.

- Krinsky, I., and A. Robb. 1990. On approximating the statistical properties of elasticities: A correction. *Review of Economics and Statistics* 72: 189–190.
- Krinsky, I., and A. L. Robb. 1986. On approximating the statistical properties of elasticities. *Review of Economics and Statistics* 68: 715–719.
- Mullahy, J. 1997. Instrumental-variable estimation of count data models: Applications to models of cigarette smoking behavior. *Review of Economics and Statistics* 79: 586–593.
- Newey, W. K., and D. McFadden. 1994. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, vol. 4, ed. R. F. Engle and D. L. McFadden, 2111–2245. Amsterdam: Elsevier.
- Terza, J. V. 2006. Estimation of policy effects using parametric nonlinear models: A contextual critique of the generalized method of moments. *Health Services and Outcomes Research Methodology* 6: 177–198.
- . 2016a. Simpler standard errors for two-stage optimization estimators. *Stata Journal* 16: 368–385.
- . 2016b. Inference using sample means of parametric nonlinear data transformations. *Health Services Research* 51: 1109–1113.
- Terza, J. V., A. Basu, and P. J. Rathouz. 2008. Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics* 27: 531–543.
- White, H. 1994. *Estimation, Inference and Specification Analysis*. New York: Cambridge University Press.
- Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA: MIT Press.
- . 2014. Quasi-maximum likelihood estimation and testing for nonlinear models with endogenous explanatory variables. *Journal of Econometrics* 182: 226–234.
- . 2015. Control function methods in applied econometrics. *Journal of Human Resources* 50: 420–445.

About the author

Joseph V. Terza is a health economist and econometrician in the Department of Economics at Indiana University–Purdue University Indianapolis. His research focuses on the development and application of methods for estimating qualitative and limited dependent variable models with endogeneity. Two of his methods have been implemented as Stata commands. He was a keynote speaker at the Stata Users Group meeting in Mexico City in November 2014.

Appendix: Stata and Mata do-files and log files for the example

Stata and Mata code

```
. /*****
> ** Read in the data.          **
> *****/
. use mullahy-birthweight-data-lbs-not-oz
.
. /*****
> ** Generate the binary smoking variable.  **
> *****/
. generate ANYCIGS=CIGSPREG>0
.
. /*****
> ** 2SRI first-stage first-part probit estimates.**
> *****/
. /*Step c*/
. probit ANYCIGS PARITY WHITE MALE EDFATHER EDMOTHER
>      FAMINCOM CIGTAX88
Iteration 0:   log likelihood = -593.2711
Iteration 1:   log likelihood = -539.2207
Iteration 2:   log likelihood = -537.93241
Iteration 3:   log likelihood = -537.9313
Iteration 4:   log likelihood = -537.9313

Probit regression               Number of obs   =       1,388
                               LR chi2(7)        =       110.68
                               Prob > chi2        =       0.0000
                               Pseudo R2         =       0.0933

Log likelihood = -537.9313
```

ANYCIGS	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
PARITY	.0183594	.0470494	0.39	0.696	-.0738558 .1105746
WHITE	.2484636	.1148504	2.16	0.031	.023361 .4735663
MALE	-.1628769	.0864755	-1.88	0.060	-.3323658 .006612
EDFATHER	-.0239095	.0100267	-2.38	0.017	-.0435614 -.0042576
EDMOTHER	-.1199751	.0216733	-5.54	0.000	-.162454 -.0774962
FAMINCOM	-.0092103	.0032144	-2.87	0.004	-.0155104 -.0029101
CIGTAX88	.0127688	.0056673	2.25	0.024	.0016611 .0238766
_cons	.5600838	.2908317	1.93	0.054	-.0099359 1.130104

```
.
. /*****
> ** Save the 2SRI first-stage first-part probit **
> ** predicted values for use in calculating      **
> ** the first stage residuals.                  **
> *****/
. /*Step d*/
. predict CIGPROB
(option pr assumed; Pr(ANYCIGS))
.
. /*****
> ** Save the first-stage first-part probit      **
> ** estimates and estimated covariance matrix.   **
> *****/
```

```

. /*Step e*/
. mata: alphas_hat=st_matrix("e(b)")
. mata: Valphas_hat=st_matrix("e(V)")
.
. /*****
> ** 2SRI first-stage second-part probit NLS      **
> ** estimates.                                **
> *****/
. /*Step c*/
. glm CIGSPREG PARITY WHITE MALE EDFATHER EDMOTHER
>      FAMINCOM CIGTAX88 if ANYCIGS==1,
>      family(gaussian) link(log) vce(robust)
Iteration 0:   log pseudolikelihood = -768.10967
Iteration 1:   log pseudolikelihood = -751.55365
Iteration 2:   log pseudolikelihood = -750.05496
Iteration 3:   log pseudolikelihood = -750.05493

Generalized linear models               No. of obs   =       212
Optimization      : ML                  Residual df   =       204
                                          Scale parameter = 71.99373
Deviance          = 14686.72175          (1/df) Deviance = 71.99373
Pearson           = 14686.72175          (1/df) Pearson  = 71.99373
Variance function: V(u) = 1              [Gaussian]
Link function     : g(u) = ln(u)         [Log]
                                          AIC           = 7.151462
Log pseudolikelihood = -750.0549266      BIC           = 13593.98

```

CIGSPREG	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
PARITY	.1004253	.0752068	1.34	0.182	-.0469773	.2478279
WHITE	.0002311	.11928	0.00	0.998	-.2335533	.2340156
MALE	.2066734	.0968097	2.13	0.033	.0169298	.396417
EDFATHER	-.0157006	.0109983	-1.43	0.153	-.0372569	.0058557
EDMOTHER	-.027413	.031649	-0.87	0.386	-.0894439	.034618
FAMINCOM	.0011098	.0039345	0.28	0.778	-.0066017	.0088212
CIGTAX88	-.0028822	.0074149	-0.39	0.697	-.0174151	.0116507
_cons	2.821627	.4702037	6.00	0.000	1.900044	3.743209

```

.
. /*****
> ** Save the 2SRI first-stage second-part NLS      **
> ** (glm) predicted values for use in calculating**
> ** the first-stage residuals.                    **
> *****/
. /*Step d*/
. predict CIGMEAN
(option mu assumed; predicted mean CIGSPREG)
.
. /*****
> ** Generate the first-stage residuals.            **
> *****/
. /*Step d*/
. generate Xuhat=CIGSPREG-CIGPROB*CIGMEAN
.

```

```

. /*****
> ** Save the first-stage second-part NLS      **
> ** estimates and estimated covariance matrix. **
> *****/
. /*Step e*/
. mata: alpha2hat=st_matrix("e(b)")
. mata: Valpha2hat=st_matrix("e(V)")

```

```

. /*****
> ** Descriptive statistics.      **
> *****/
. summ

```

Variable	Obs	Mean	Std. Dev.	Min	Max
BIRTHWT	1,388	118.6996	20.35396	23	271
CIGSPREG	1,388	2.087176	5.972688	0	50
PARITY	1,388	1.632565	.8940273	1	6
WHITE	1,388	.7845821	.4112601	0	1
MALE	1,388	.5208934	.4997433	0	1
EDFATHER	1,388	11.32421	5.251299	0	18
EDMOTHER	1,388	12.92651	2.401109	0	18
FAMINCOM	1,388	29.02666	18.73928	.5	65
CIGTAX88	1,388	19.55295	7.795598	2	38
BIRTHWTLB	1,388	7.418723	1.272123	1.4375	16.9375
ANYCIGS	1,388	.1527378	.3598642	0	1
CIGPROB	1,388	.1520482	.1038465	.0049521	.7636681
CIGMEAN	1,388	12.86834	2.512522	7.946904	28.78438
Xuhat	1,388	.0063805	5.818791	-15.09198	46.96746

```

.

```

```

. /*****
> ** 2SRI second-stage NLS estimates.      **
> *****/
. /*Step f*/
. glm BIRTHWTLB CIGSPREG PARITY WHITE MALE Xuhat,
>      family(gaussian) link(log) vce(robust)

Iteration 0:  log pseudolikelihood = -2271.1401
Iteration 1:  log pseudolikelihood = -2263.4591
Iteration 2:  log pseudolikelihood = -2263.4109
Iteration 3:  log pseudolikelihood = -2263.4109

Generalized linear models              No. of obs      =       1,388
Optimization      : ML                 Residual df      =       1,382
                                         Scale parameter =    1.533962
Deviance          =    2119.935722      (1/df) Deviance =    1.533962
Pearson           =    2119.935722      (1/df) Pearson =    1.533962
Variance function: V(u) = 1             [Gaussian]
Link function     : g(u) = ln(u)        [Log]
                                         AIC              =    3.270045
                                         BIC              =   -7879.69

Log pseudolikelihood = -2263.410887

```

BIRTHWTLB	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
CIGSPREG	-.0119672	.0027167	-4.41	0.000	-.0172918	-.0066427
PARITY	.0183912	.0050259	3.66	0.000	.0085405	.0282419
WHITE	.0542038	.0117566	4.61	0.000	.0311614	.0772463
MALE	.0259255	.0089519	2.90	0.004	.0083802	.0434708
Xuhat	.0077064	.0026665	2.89	0.004	.00248	.0129327
_cons	1.942015	.0149736	129.70	0.000	1.912667	1.971363

```

.
. /*****
> ** Save second-stage estimates and covariance **
> ** matrix. Single out the coefficient estimate **
> ** for Xu.      **
> *****/
. /*Step g*/
. mata: betahat=st_matrix("e(b)")
. mata: Vbetahat=st_matrix("e(V)")
. mata: Bu=betahat[5]
.
. /*****
> ** Send the requisite variables to Mata as      **
> ** vectors.      **
> *****/
. /*Step h*/
. putmata BIRTHWTLB CIGSPREG ANYCIGS PARITY WHITE
>      MALE EDFATHER EDMOTHER FAMINCOM CIGTAX88 Xuhat
(11 vectors posted)
.

```

```

. /*****
> ** Use these vectors to concatenate the needed **
> ** matrices. **
> *****/
. /*Step h*/
. mata: X=CIGSPREG, PARITY, WHITE, MALE,
>       Xuhat, J(rows(PARITY),1,1)
. mata: W=PARITY, WHITE, MALE, EDFATHER, EDMOTHER,
>       FAMINCOM, CIGTAX88, J(rows(PARITY),1,1)
.
. /*****
> ** Set up the two gradient matrices for the ACSE**
> *****/
. /*Step i*/
. mata: gradbeta=exp(X*betahat):*X
. mata: gradalpha1=-Bu:*exp(X*betahat):*normalden(W*alpha1hat):*exp(W*alpha2hat):*W
. mata: gradalpha2=-Bu:*exp(X*betahat):*normal(W*alpha1hat):*exp(W*alpha2hat):*W
. mata: gradalpha=gradalpha1,gradalpha2
.
. /*****
> ** Set up the B1 and B2 matrices for the ACSE. **
> *****/
. /*Step j*/
. mata: B1=gradbeta'*gradbeta
. mata: B2=gradbeta'*gradalpha
.
. /*****
> ** Set up the full estimated asymptotic **
> ** covariance matrix for alpha (first-stage **
> ** two-part model covariance matrix estimator as**
> ** output by Stata). **
> *****/
. /*Step k*/
. mata: Valphahat=blockdiag(Valpha1hat,Valpha2hat)
.
. /*****
> ** Construct the covariance matrix of the **
> ** second-stage Beta estimates. **
> *****/
. /*Step k*/
. mata: Dhat=invsym(B1)*B2*Valphahat*B2'*invsym(B1)+Vbetahat
.
. /*****
> ** Extract the vector of asymptotically correct **
> ** standard errors for betahat. **
> *****/
. /*Step l*/
. mata: ACSE=sqrt(diagonal(Dhat))
.

```

```

. /*****
> ** Calculate the corresponding vector of      **
> ** asymptotically correct t-stats.          **
> *****/
. /*Step m*/
. mata: ACtstats=betahat:/ACSE

. /*****
> ** Compute the corresponding vector of p-values.
> *****/
. mata: ACpvalues=2*(1:-normal(abs(ACtstats)))

. /*****
> ** Display results.                          **
> *****/
. mata: header="Variable","Estimate","ACSE","AC t-stat","pvalue"
. mata: varnames="CIGSPREG", "PARITY", "WHITE", "MALE", "Xuhat","Constant"
. mata: results=betahat,ACSE,ACtstats,ACpvalues
. mata: resview=strofreal(results)
. mata: "2SRI Results with ACSE"
      2SRI Results with ACSE
. mata: header \ (varnames`,resview)

```

	1	2	3	4	5
1	Variable	Estimate	ACSE	AC t-stat	pvalue
2	CIGSPREG	-.0119672	.002939	-4.071839	.0000466
3	PARITY	.0183912	.0054684	3.363166	.0007705
4	WHITE	.0542038	.0121787	4.450694	8.56e-06
5	MALE	.0259255	.009266	2.797918	.0051433
6	Xuhat	.0077064	.0028991	2.658169	.0078566
7	Constant	1.942015	.0155771	124.6715	0